Network Infrastructure Visualisation Using High-Dimensional Node-Attribute Data

Helen Gibson* Paul Vickers†

Northumbria University

ABSTRACT

We present an extended version of targeted projection pursuit, a high dimensional data exploration tool adapted for producing graph layouts using node-attributes. Attributes are generated based on detected events in the intrusion detection system and firewall logs and how often they occur for each IP address. Edges are the directed links between source and destination IPs. The layout is interactive and users can manipulate the points in order to find interesting layouts and then further analyse how these layouts are related to the events in the logs. Thus, they first allow the user to detect anomalies and then gives them a platform to investigate why they occur.

1 INTRODUCTION

Mini challenge two for the VAST challenge 2012 was to identify why a number of users in a network were getting virus alerts and messages from unknown anti-virus systems. The setting was a particular regional office in the Bank of Money (BOM) corporate network which now operates 24-hours a day to support its function as a call-center. The task, through the use of firewall and intrusion detection system logs, is to identify the noteworthy events in the system, any security trends, the root causes of the problems and, following this, to recommend actions to prevent this happening again in the future. The approach taken has been to adapt this data into a node-attribute graph and to use an extended version of targeted projection pursuit (TPP) [1, 2] to analyse this data as a graph.

2 DATA PREPARATION

In order to use this data in TPP it was converted into a node-attribute style format. By identifying each priority, classification and label in the IDS logs, SQL queries against the raw data were then run in R [4] that counted for each IP address in the system the number of times it had been detected as being involved in a particular event. This was done for both days individually and then as a combined IDS data set. These queries were then run a second time but the number of attributes was doubled so as to classify whether the event happened when the IP address was acting as a source node or as a destination node. For the firewall logs a different approach was taken and the attributes became 15 minute intervals in the data logs and the number of events each IP address was involved in in that period became the value of the attribute. Again this was also redone to reclassify the time periods as being associated with source or destination nodes. These files were then loaded into Weka [3] in order to normalise the data on a scale between 0 and 1 (thus not punishing events that only occur a few times) and to output it in ARFF format for use with TPP. From the data provided IP addresses were then given a classification such as workstation, Bank of Money accessible website, DNS, etc. In this classification we were able to identify our first irregularity in the firewall dataset - a number of nodes with IP address beginning 172.28.29.X which are not known to be part of the network.

*e-mail: helen.gibson@northumbria.ac.uk

†e-mail:paul.vickers@northumbria.ac.uk

IEEE Conference on Visual Analytics Science and Technology 2012
October 14 - 19, Seattle, WA, USA
978-1-4673-4753-2/12/$31.00 ©2012 IEEE

3 EXTENDING TPP

Targeted projection pursuit [1] (TPP) was originally developed as a tool to explore high-dimensional data sets interactively with a particular focus on being able to cluster the data. The basic idea being that the user begins with a two-dimensional principal component analysis projection of a high-dimensional data set. The user is then encouraged to grab points and move them around, attempt to cluster them or do any other manipulations that they believe could be interesting. As the points are moved TPP searches to try to find a projection from the high-dimensional data to the one requested by the user on screen and displays the closest possible projection it can find. Users can then use a table on the right-hand side of the tool to identify which attributes are contributing most significantly to the projection. An extension to this where the user is able to show a set of edges in a graph, thus positioning the nodes in the projection becomes a graph layout method, has already been published [2].

Since then TPP has been extended in a number of ways. Users can elect to colour the edges according to their source node in order to indicate direction, they may also elect to show arrows or to curve the edges clockwise also to show direction. Particularly in large datasets items can become obscured or a user may wish to focus on a particular part of the graph and so they can pan and zoom around the graph. To focus on the graph part of the layout when a user selects a number of nodes their edges are shown fully opaque while other edges in the graph become transparent. A edge filter can also be turned on so that only the edges of the selected nodes are shown, which is useful in graphs with many edges and for large datasets to reduce the time spent updating the display. The labels (in this case the IP address) can be shown by hovering over a node or by electing to show labels associated with selected nodes. Specifically for this task the graph can also be connected to a MySQL database holding the raw data. Once connected if a user selects a number of nodes she can then choose to view all of the data associated with the selected IP addresses.

4 ANALYSIS

The graph in Figure 1 shows the IDS graph for both days loaded into TPP. This has been taken from its initial layout and the nodes have been rearranged to see an interesting projection. There are the BOM accessible websites located along one axes from left to right and just below that most of the workstations along a second axes running left to right. In the top left corner is the node representing the firewall interface to the regional bank network and five workstations whose attributes are different to the majority of the other workstations located nearby. The links of these five nodes can clearly be seen. Selecting these nodes allows the user to see how the values of their attributes compare to the rest of the graph. It is found that they are involved in events to do with accessing email and database ports as well as SNMP requests and VNC scans. Bringing up the raw data associated with these nodes through the database shows that there seems to be an ordered pattern to these events.

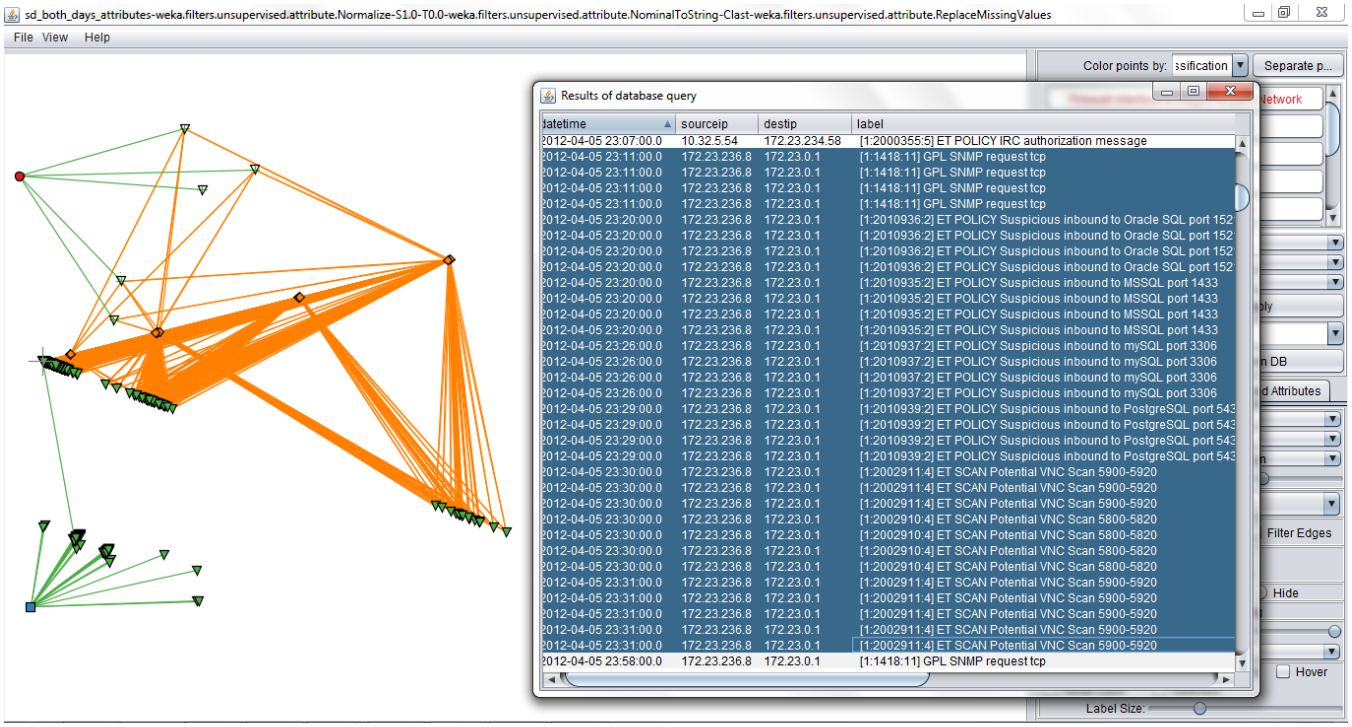

Figure 1: The IDS graph. The orange diamonds represent BOM accessible websites, the green triangles workstations, the red circle the firewall interface to the regional bank network and the blue square the DNS. The database query window shows the raw data for five selected workstations (the crossed triangles) and we see a pattern of events in the label column.

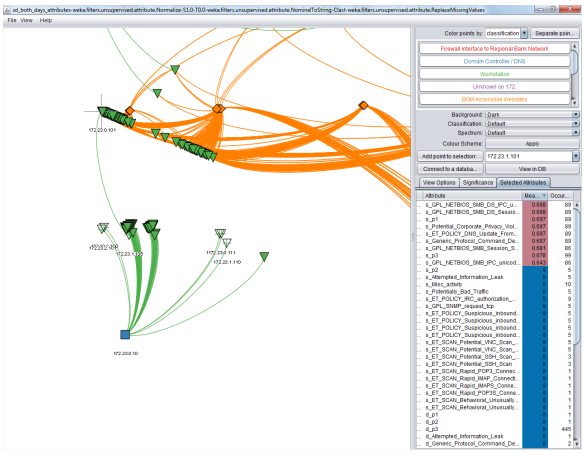

Figure 2: Zoomed in portion of the IDS graph to those nodes connecting to the DNS. The right hand table shows that these nodes have higher than the average number of events related to NETBIOS.

In the bottom left hand corner are the workstations that connect to the DNS. These also have a different set of attributes to the rest of the workstation nodes, something which can be seen more clearly in Figure 2. The right hand panel shows how these are all events related to the NETBIOS which can then be investigated further in the database. Otherwise a user may decide to pay more attention to how these nodes link to the DNS such as if there is any significance in there being two clusters of nodes and then a few outliers? Or why there appears to be just one link between two workstations out of all of the edges in the graph and if it is something worth investigating further?

5 Conclusion

It has been shown how TPP can be adapted for use as an exploratory graph visualisation and analysis tool in an attempt to visualise vulnerabilities in a corporate call-center network. TPP has allowed for the immediate identification of outliers, and therefore potential security issues in just a few steps. It has also demonstrated how edge-attribute data can be manipulated into node-attribute data in order to explore graphs in this context. It has been able to create a large graph visualisation without resulting in a hairball layout but instead one that can be informative and stimulate further analysis and exploration of the graph. A potential improvement to this method would be to incorporate edge weights into the visualisation since both one unusual connection or an extreme number of connections could both prove significant.

References

[1] J. Faith. Targeted projection pursuit for interactive exploration of high-dimensional data sets. In *11th International Conference Information Visualization (IV 07)*, pages 286–292, Jul 2007.

[2] H. Gibson and J. Faith. Node-attribute graph layout for small-world networks. In *15th International Conference on Information Visualisation*, pages 482–487, Jul 2011.

[3] M. Hal, E. Frank, G. Holmes, B. Bernhard Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, 11, 2009.

[4] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.